02

DATA PREPROCESSING

CHAPTER - 02

WELCOME TO THE NEW WAY OF LEARNING!

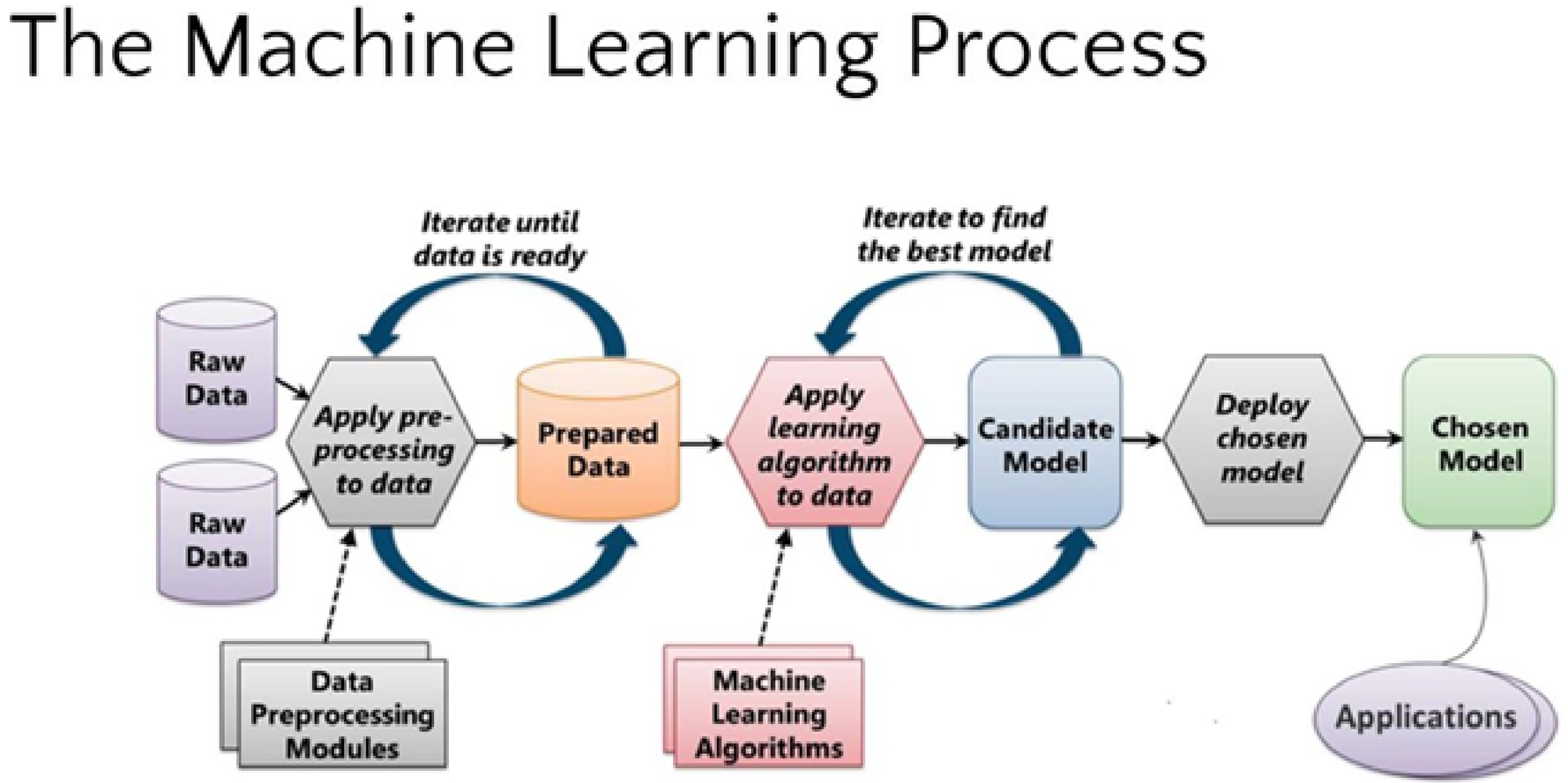
Data Pre processing

- Technique that involves **transforming** raw **data** into an understandable format

- Transforming the data such that it becomes **machine-readable**

- The *features* of the data can now be *easily* interpreted by the **algorithm**.

Data Pre processing



Why Data Preprocessing?

- **Data in the real world is dirty**
  - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - **noisy:** containing errors or outliers
  - **inconsistent:** containing discrepancies incodes or names

- **No quality data, no quality results!**
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Why Data Preprocessing?

- **A multi-dimensional measure of data quality:**
  - A well-accepted multi-dimensional view:
  - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility

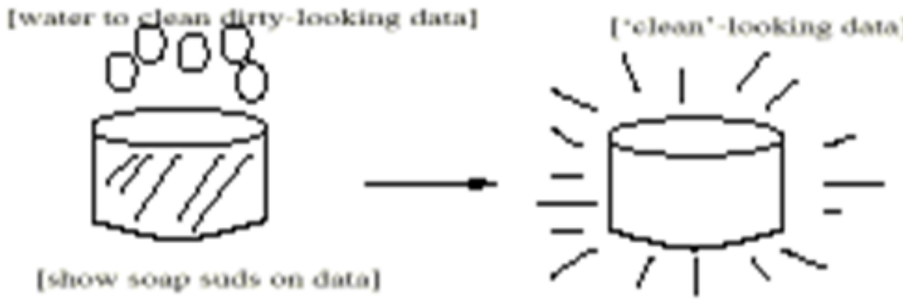
- **Broad categories:**
  - lintrinsic, contextual, representational, and accessibility.

Major Tasks in Data Preprocessing?

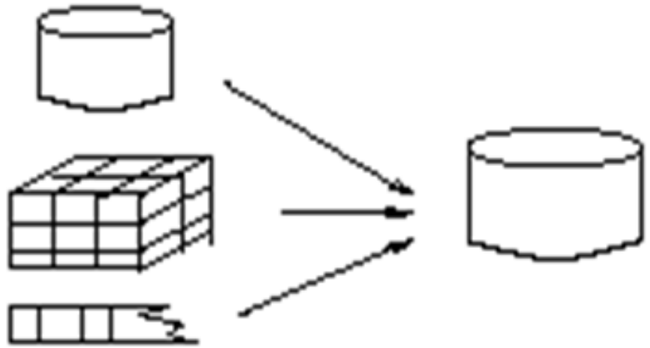
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, files, or notes
- **Data transformation**
  - Normalization (scaling to a specific range)
  - Aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - **Data discretization:** with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression, generalization

Forms of Data Preprocessing?

Data Cleaning

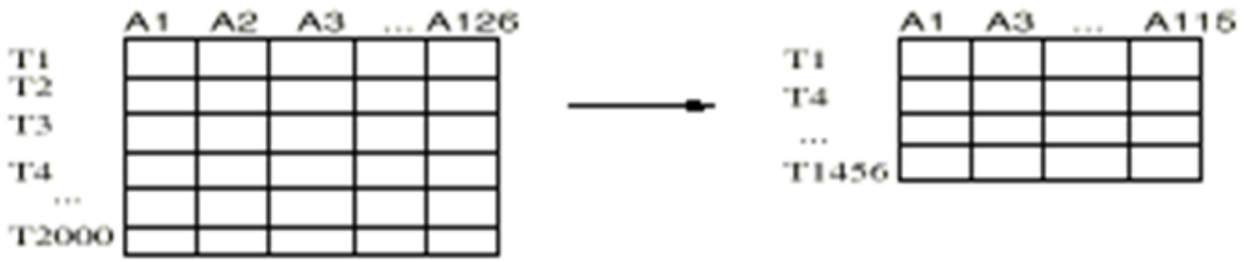


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

- **Data cleaning tasks**
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- **Data is not always available**
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- **Missing data may be due to**
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- **Missing data may need to be inferred**

How to Handle Missing Data?

- **Ignoring the tuple:** usually done when class label is missing (assuming the task is classification — not effective in certain cases)
- Fill in the missing value **manually**: **tedious + infeasible**
- Use a global constant to fill in the missing value: e.g., — unknown‖, a new class?!
- Use the **attribute mean** to fill in the **missing** value
- Use the attribute mean for all samples of the same class to fill in the missing value: **smarter**
- Use the most probable value
    - **inference-based such as regression, Bayesian formula, decision-tree**

Noisy Data

Q: What is noise?

A: Random error in a measured variable.

- **Incorrect attribute values may be due to**
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems which require data cleaning**
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning method:**
  - first **sort** data and **partition** into (equi-depth) **bins**
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for **discretization**
- **Clustering**
  - **detect** and **remove outliers**
- **Semi-automated method: combined computer and human inspection**
  - detect suspicious values and check manually
- **Regression**
  - smooth by **fitting** the **data** into **regression functions**

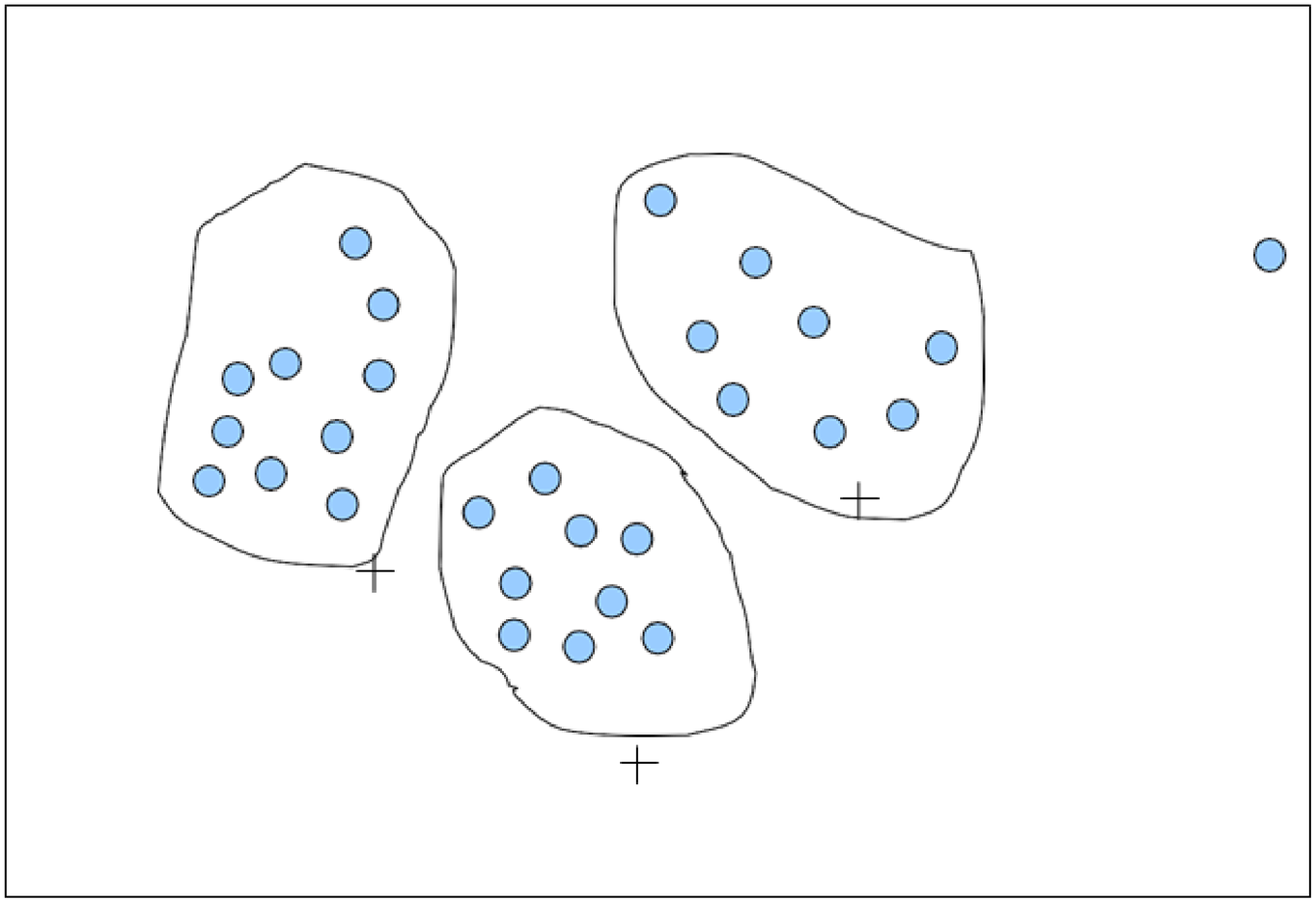
Simple Discretization Methods: Binning

- **Equal-width (distance) partitioning:**
  - It divides the range into **N** intervals of equal size: **uniform grid**
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = **(B-A)/N**
  - The most straight forward
  - But outliers may dominate presentation
  - Skewed data is not handled well.
- **Equal-depth (frequency) partitioning:**
  - It divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

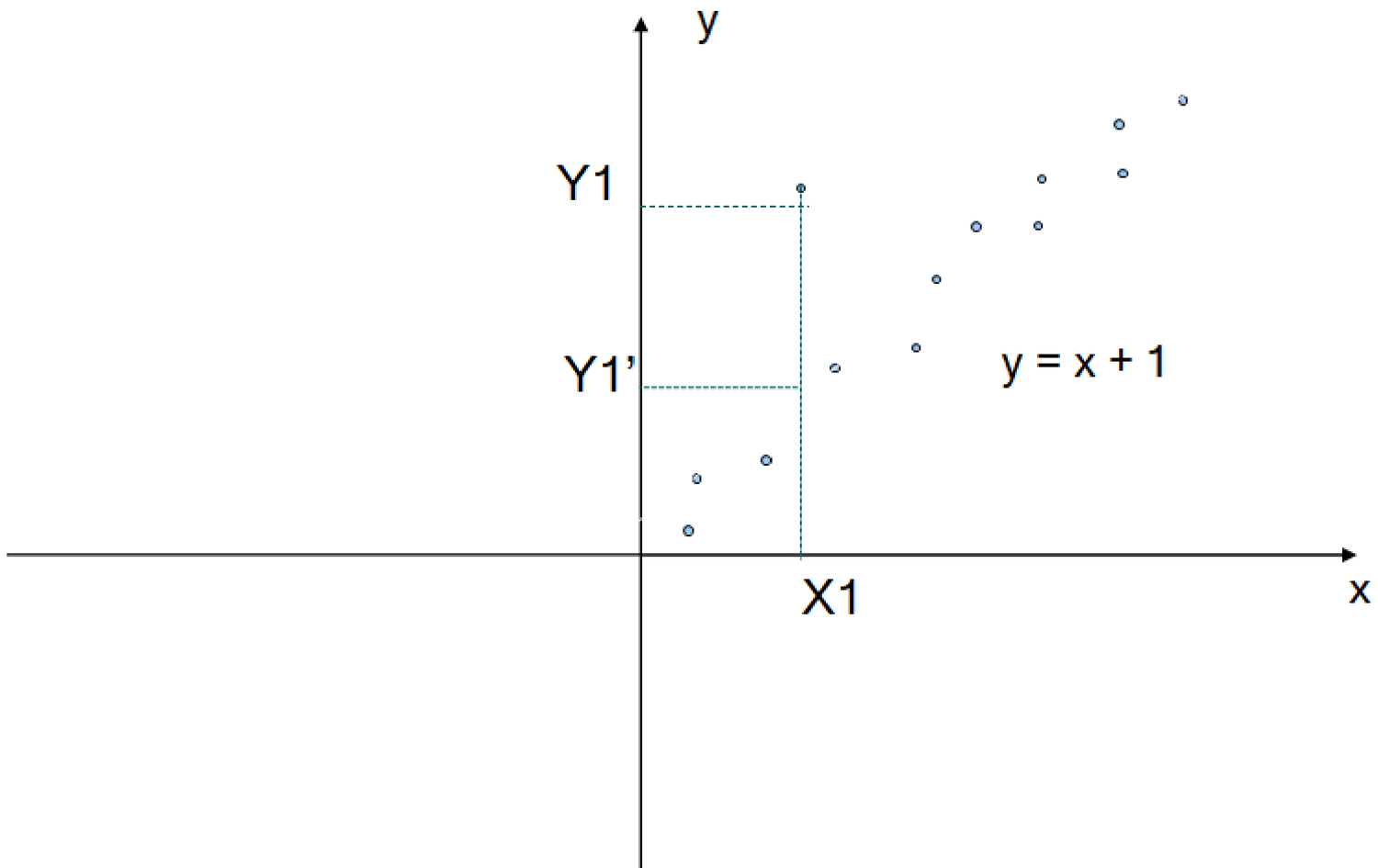
- **Sorted data for price (in dollars)**: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- **Partition into (equi-depth) bins:**
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
- **Smoothing by bin means:**
    - Bin 1: 9, 9, 9,9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
- **Smoothing by bin boundaries:**
    - Bin 1: 4, 4, 4, 15
    - Bin 2: 21, 21, 25, 25
    - Bin 3: 26, 26, 26, 34

Cluster Analysis



Regression

- **Linear regression** (best line to fit two variables)
- **Multiple linear regression** (more than two variables, fit to a multidimensional surface)



How to Handle Inconsistent Data?

- **Manual correction using external references**
- **Semi-automatic using various tools**
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Data Integration

- **Data integration:**
  - combines data from multiple sources into a coherent store
- **Schema integration**
  - integrate metadata from different sources
  - Entity identification problem: identify real-world entities from multiple data sources, **e.g., A.cust-id == B.cust-#**
- **Detecting and resolving data value conflicts**
  - for the same real-world entity, attribute values from different sources are different
  - **possible reasons:** different representations, different scales, e.g., metric vs. British units, different currency

Handling Redundant Data in Data Integration

- **Redundant data occur often when integrating multiple DBs**
  - The same attribute may have different names in different databases
  - One attribute may be a — derived || attribute in another table, e.g., annual revenue

- Redundant data may be able to be detected by **correlational analysis**

- **Careful integration** help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- **Smoothing:** remove noise from data (binning, clustering, regression)
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction**
  - New attributes constructed from the given ones

Data Reduction

- **Problem:**

  Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- **Solution?**

  Data reduction…

Data Reduction

- Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

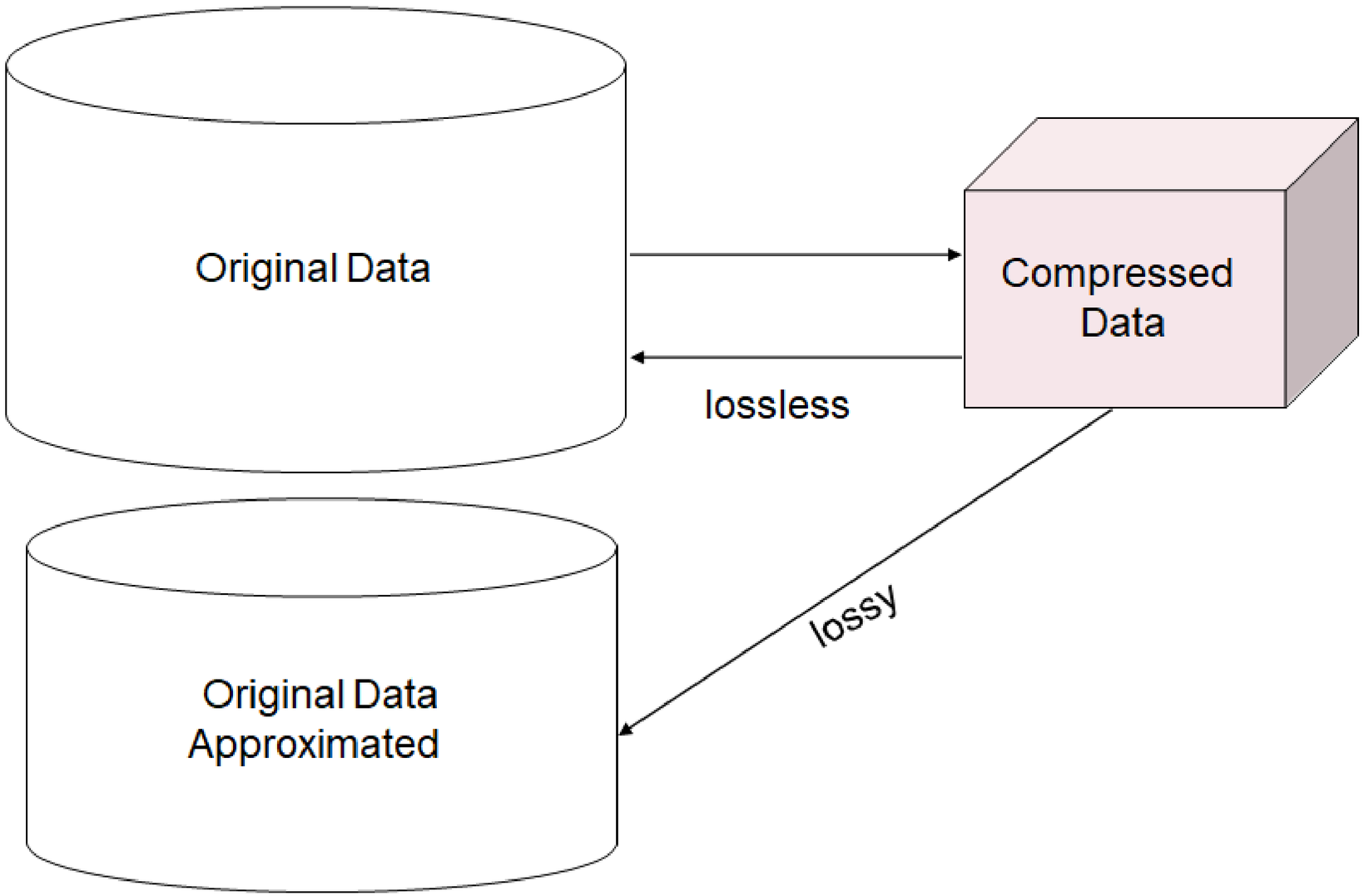
- **Data reduction strategies**
  - Dimensionality reduction
  - Datacompression
  - Numerosity reduction

Dimensionality Reduction

- **Problem**: Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)

- **Solution**: Heuristic methods (due to exponential # of choices) usually greedy:
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction

Data Compression



Numerosity Reduction

- **Parametric methods**
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - E.g.: Log-linear models: obtain value at a point in m- D space as the product on appropriate marginal subspaces

- **Non-parametric methods**
  - Do not assume models
  - Major families: histograms, clustering, sampling

Regression and Log-Linear Models

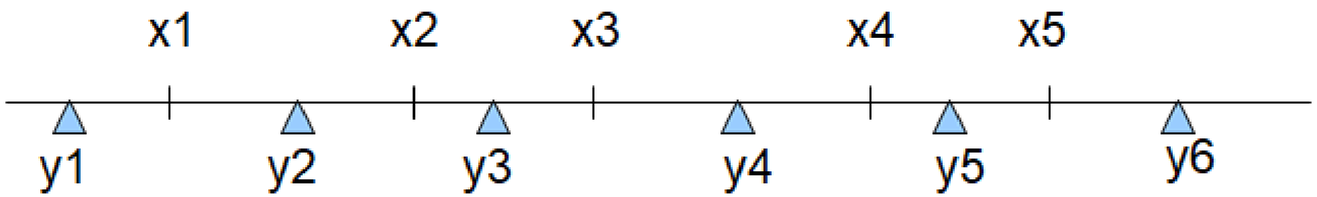
- **Linear regression:** Data are modelled to fit a straight-line:
    - Often uses the least-square method to fit the line

- **Multiple regression:** allows a response variable y to be modelled as a linear function of multidimensional feature vector (predictor variables)

- **Log-linear model:** approximates discrete multidimensional joint probability distributions

Discretization/Quantization

- **Three types of attributes:**
  - **Nominal** — values from an unordered set
  - **Ordinal** — values from an ordered set
  - **Continuous** — real numbers
- **Discretization/Quantization:**
  - divide the range of a continuous attribute into intervals



- Some classification algorithms only accept categorical attributes.
- Reduce data size bydiscretization

Summary

- **Data preparation is a big issue for both warehousing and mining**

- **Data preparation includes**
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization

- **A lot of methods have been developed but still an active area of research**


NINJA
Co.
Thank You !

KNOWLEDGE WILL GIVE YOU POWER, BUT CHARACTER RESPECT.
– BRUCE LEE

THE SUCCESS OF OUR VIRTUAL CLASSROOM LIES IN OUR COLLECTIVE HANDS!

QUESTIONS? COMMENTS?

Feel free to share your feedback.

EMAIL ADDRESS

ninjaprogrammercorp@gmail.com

MOBILE NUMBER

123-456-7890

CONSULTATION HOURS

4 PM to 6 PM